More Than Words: Inference of Socially Relevant Information From Nonverbal Vocal Cues in Speech

A.Vinciarelli[1,2], H.Salamin[1], G.Mohammadi[2,3], and K.Truong[4]

[1] University of Glasgow, Sir A.Williams Bldg., G12 8QQ Glasgow (UK)
{vincia,hsalamin}@dcs.gla.ac.uk
[2] Idiap Research Institute, CP592, 1920 Martigny (Switzerland)
gmohamma@idiap.ch
[3] EPFL, 1015 Lausanne (Switzerland)
[4] University of Twente, Drienerlolaan 5, Enschede (The Netherlands)
k.p.truong@ewi.utwente.nl

Abstract. This paper presents two examples of how nonverbal communication can be automatically detected and interpreted in terms of social phenomena. In particular, the presented approaches use simple prosodic features to distinguish between journalists and non-journalists in media, and extract social networks from turn-taking to recognize roles in different interaction settings (broadcast data and meetings). Furthermore, the article outlines some of the most interesting perspectives in this line of research.

Keywords: Social Signal Processing, Turn-Taking, Prosody, Nonverbal Behavior, Roles, Personality, Speaking Style

1 Introduction

There is more than words in human-human interaction. Even if our attention focuses on the verbal content of the messages being exchanged (what people say), we still perceive and interpret the wide spectrum of nonverbal behavioral cues that people display when they talk with others like facial expressions, vocalizations, gestures, postures, etc. These are the signals that help us to understand, beyond the face value of the words we listen to, affective, emotional and social aspects of the interactions we are involved in [14].

Nonverbal communication has been studied mainly by human sciences (psychology, anthropology, sociology, etc.), but recently it has attracted significant attention in the computing community as well. The reason is that nonverbal behavioral cues like those mentioned above are the physical, machine detectable evidence of phenomena non-otherwise observable such as emotions, social attitudes, intentions, etc. This means that the cues, when detected through sensors and interpreted through machine intelligence approaches, can help machines to become socially and emotionally intelligent, i.e. capable of dealing with human-human interactions like humans do.

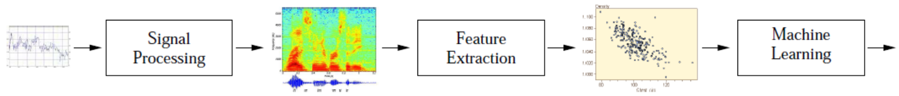

Fig. 1. Journalist / non-journalist recognition approach.

This work focuses on two kinds of cues, those related to turn-taking (who talks when, to whom, and how much in a conversation) and the speaking style (the way a person talks). While being only two of the myriads of cues that humans can display, turn-taking and speaking style are important because they are crucial in conversations, the most common and primordial site of human-human interaction.

In particular, this work shows how the turn-taking can be analyzed and modeled to recognize automatically the roles that people play in different interaction settings (broadcast data and meetings), and how the speaking style can be used to discriminate between journalists and non-journalists in broadcast data. In both cases, detection and automatic interpretation of nonverbal cues is shown to be an effective means to extract high level information from spoken data, in accordance to indications coming from human sciences.

The rest of the paper is organized as follows: Section 2 shows how the speaking style heps to distinguish between journalists and non-journalists, Section 3 presents an approach for automatic role recognition based on turn-taking and Social Networks, and the final Section 4 draws some conclusions.

2 Prosody: Spotting Journalists in Broadcast Data

When listening to radio news, it is usually easy to tell whether a speaker is a journalist or not. As journalists are professional speakers who know how to manipulate their voices to keep attention, it is not surprising that their speaking style is peculiar. This difference can come from numerous features, like intonation, pauses, rhythm, harmony and even choice of words. It is still not clear which cues enable the listeners to make a distinction between different styles. Studying on different types of speaking style helps to better understand the process of speech production and perception [4, 9, 15]. Therefore, areas like speech synthesis, speech recognition and verbal behavior analysis can benefit from speaking style analysis.

There is no taxonomy or dictionary of speaking styles and researchers in the field typically define and compare ad-hoc styles pertinent to their studies, for example spontaneous vs. read speech, slow vs. fast, or stressed vs. non-stressed [10]. The experiments performed in this work focus on the distinction between journalists and non-journalists that can be considered a particular case of the most general distinction between professional and non-professional speaking.

The block diagram of our approach is illustrated in Figure 1. The first stage extracts short-term prosodic features, namely pitch, formants, energy and rhythm: The pitch is the oscillation frequency of the vocal folds, the formants are the frequencies corresponding to the resonance of the vocal tract, the energy is the amplitude of the speech signal, and the rythm is estimated indirectly through the length of voiced and unvoiced segments, the faster the speech rate, the shorter, on average, the segments. All of these features are extracted from 40 ms analysis windows at regular time steps of 10 ms using Praat, one of the most commonly applied speech analysis tools [5].

All of the above features account for short term phenomena because they are extracted from short analysis windows, but the speaking style is a longer term property of speech. Thus, the above features are not used directly, but through functionals that account for their statistical properties. The goal of the second stage in the scheme of Figure 1 is exactly to estimate these properties. In this work, this is done through the entropy of the short-term features. If f is one of the short-term features mentioned above, the entropy is estimated as follows:

$$H(f) = \frac{\sum_{i=1}^{|F|} p(f_i) \log p(f_i)}{\log |F|} \tag{1}$$

where $F = \{f_1, \ldots, f_{|F|}\}$ is the set of f values during an interval of time, and $|F|$ is the cardinality of F. The long-term features are expected to capture the variability of each short-term feature, the higher the entropy, the higher the number of f values represented a large number of times during a long time interval and viceversa.

The third stage of the approach is the classification of the speaking style, represented with a vector including the entropy of the six short term features mentioned at the beginning of this section. The classification is performed with a Support Vector Machine with a Radial Basis Function kernel, an algorithm that uses a set of training samples to identify a discriminant hyperplane expected to separate feature vectors corresponding to different classes.

In the experiments of this work, the SVM is trained with a k-fold approach: The entire dataset is split into k equal size subsets, and $k-1$ parts are used for training the model while the remaining part for testing. This procedure is repeated k times, (each time, one of the subsets is used for testing) and the average error of all k runs will be reported as classification performance measure [2, 7], in the experiments of this work, $k = 10$.

The experiments have been performed over a corpus of 686 audio clips including 313 non-journalists and 373 journalists, for a total of 330 identities. The average percentage of clips classified correctly is 88.4. The recognition rate for journalist and non-journalist is 87.3 and 88.4 respectively. In this experiment we have used the total length of all clips which changes for each sample.

In another experiment, we took an interval of equal length from each clip and we analysed the performances for different lengths. Figure 3 shows how the performance changes when the length of the clips increases. The plot shows that the longer clips are better classified. This is not surprising as longer clips allow a better estimation of the entropies used as features. Some misclassifications are due to the intrinsic ambiguity of the data: some non-journalist speakers, e.g.

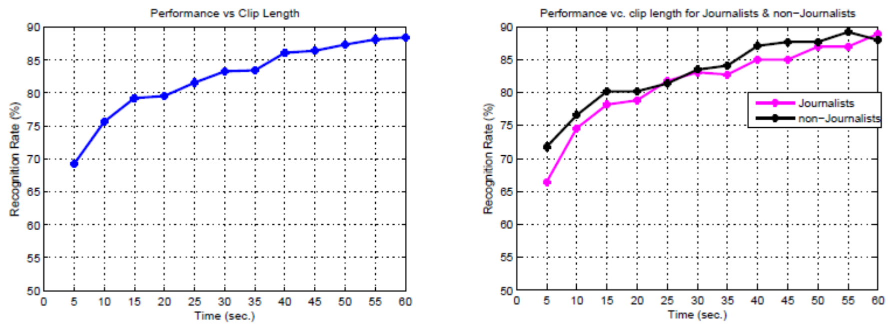

Fig. 2. Recognition performance as a function of the clips length. The right plot shows the results for the two classes separately.

politicians and actors that often appear in the media, have the same speaking proficiency as the journalists and, at the same time, some journalists are not as effective as their colleagues in delivering their message. To our knowledge, no other systems performing a similar task have been presented in the literature.

This makes it difficult to say whether the performance of the system is satisfactory or not. For this reason, the performance of the automatic system has been compared with the results obtained by 16 human assessors on a similar task. A set of 30 audio clips were randomly selected from the data corpus. In the set, 17 clips correspond to journalists and 13 to non-journalists. The length of the clips ranges from 3.5 to 75 sec. and it reproduces roughly the length distribution of the data corpus.

The human assessors have listened to the clips and have assigned each one of them to one of the two classes. In order to reduce as much as possible the influence of the content, the assessors do not speak the language of the clips (French), and their mother tongues include English (2 persons), Hindi (5 persons), Chinese (6 persons), Farsi (1 person), Serbian (1 person) and Arabic (1 person). The group of assessors includes 5 women and 11 men.

The total number of judgements made by the assessors is 480 and their overall performance, i.e. the fraction of correct judgements, is 82.3 percent. The women have an overall performance of 88 percent (on average 26.4 correct judgements out of 30), while the men have an overall performance of 79.0 percent (on average 23.7 correct judgements out of 30). On average, each clip has been recognized correctly by 13.2 assessors, but there are two ambiguous clips, recognized by only 2 and 4 assessors respectively, that reduce significantly the average. Without taking into account such clips, the average number of correct classifications per clip is 13.9.

The performance of the automatic system over the same clips submitted to the human assessors is, in the limits of the statistical fluctuations, the same. Furthermore, the system and the human assessors tend to make the same decision about the same clip. This seems to suggest that the features proposed in this work

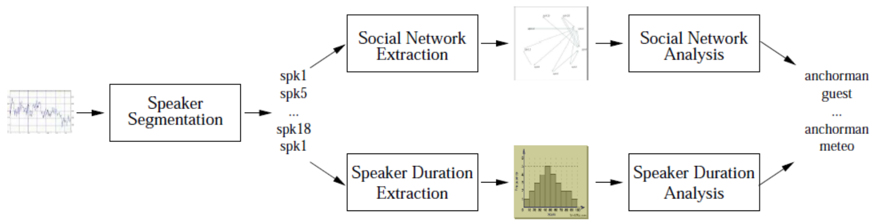

Fig. 3. Role recognition approach.

actually capture, at least in part, perceptually important aspects of nonverbal vocal behavior, but the dataset is too small for reaching definitive conclusions about this point. Unfortunately, it was not possible to ask the assessors to listen to the whole dataset (more than 7 hours in total) for practical reasons.

3 Turn-Taking: Automatic Role Recognition

Whenever they interact, people play *roles*, i.e. they display predictable behavioral patterns perceived by others as addressing interaction needs fulfilling group functions. This section shows how this phenomenon concerns one of the most salient characteristics of a conversation, namely who talks when, how much and with whom: in a single expression, the turn-taking.

To work on the turn-taking is particularly appealing from a technological point of view because there are many effective techniques for automatically segmenting audio recordings into turns. In general, these give as output a set of triples including a speaker label (a code identifying one of the speakers involved in the conversation), a start time and a duration:

$$S = \{(s_1, t_1, \Delta t_1), \dots, (s_N, t_N, \Delta t_N)\} \quad (2)$$

where N is the total number of turns and $s_i \in A = \{a_1, \dots, a_G\}$ (G is the total number of speakers in the conversation and the a_i are the speaker labels). Even if such an information is relatively basic and it seems to miss the richness of a conversation, still it allows one to capture a wide range of social phenomena such as the groups forming around discussion topics [13], the fronts opposing one another in competitive discussions [12], dominant individuals [8], etc. The rest of this section shows how the same information can be used to infer the roles in several interaction settings.

The overall approach is depicted in Figure 3 showing the different steps of the process: first the audio data is split into turns using the speaker clustering approach described in [1], then the turn-taking S is used to extract a Social Affiliation Network (see below for more details) and analyze the duration distribution of the turns. At this point, each person participating in the conversation can be represented with feature vectors and these are mapped into roles using Bayesian classifiers based on discrete distributions.

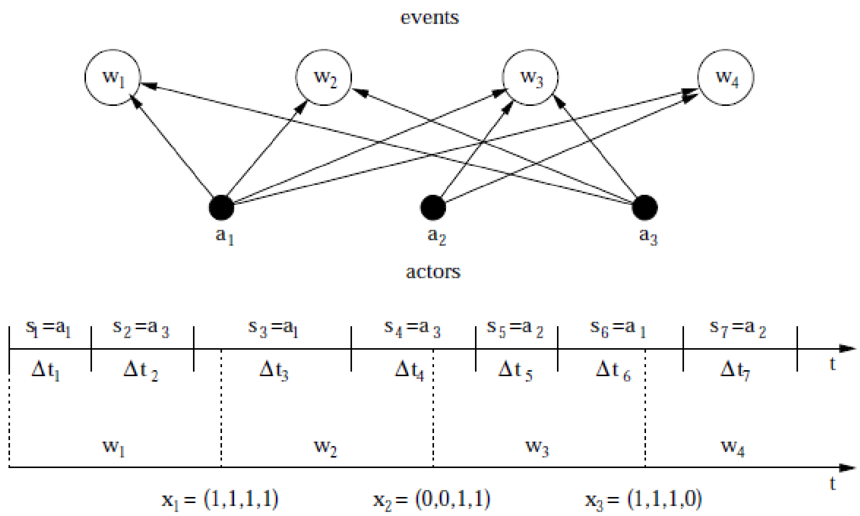

Fig. 4. Social Network Extraction.

3.1 Social Network Extraction

The turn-taking S can be used to extract a Social Affiliation Network (SAN) capturing the interaction pattern between people playing different roles. A SAN is a bipartite graph with two kinds of nodes, the *events* and the *actors*. The latter correspond to the people involved in the conversation and the former are defined following the temporal proximity principle: people talking during the same interval of time are likely to interact with one another (see below). Only nodes of different kind can be linked.

Thus, each recording is split into a number D of uniform, non-overlapping intervals w_j that are used as events. Actors are said to participate in event w_j if they talk during it. When an actor a_i participates in an event w_j the corresponding nodes are linked. The main advantage of this represebtation (commonly applied by sociologists to identify social groups) is that each actor can be represented with a tuple $\boldsymbol{x} = (x_1, \ldots, x_D)$, where $x_i = 1$ if the actor participates in the event and 0 otherwise.

The number D of intervals is a hyperparameter of the approach and must be defined via crossvalidation (see below for more details).

3.2 Role Recognition

In mathematical terms, the role recognition step corresponds to finding a mapping $\varphi : A \rightarrow \mathcal{R}$, where A is a set of actors (see above) and $\mathcal{R}$ is a set of roles, such that $\varphi(a)$ is the role of actor a. Each actor is represented with a pair $\boldsymbol{y}_a = (\tau_a, \boldsymbol{x}_a)$, where τ_a is the fraction of time during which a talks during a conversation and $\boldsymbol{x}_a$ is the tuple extracted from the Social Affiliation Network described above.

At this point, the role recognition problem can be thought of as finding the mapping $\hat{\varphi}$ such that:

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} p(Y|\varphi)p(\varphi). \tag{3}$$

where $Y = \{\mathbf{y}_a\}_{a \in A}$ is the set of pairs $\boldsymbol{y}$ corresponding to the actors of a given conversation and $\mathcal{R}^A$ is the set of all possible functions mapping actors into roles.

The problem can be simplified by making two assumptions, the first is that the observations are mutually conditionally independent given the roles. The second is that the observation $\mathbf{y}_a$ of actor a only depends on its role $\varphi(a)$ and not on the role of the other actors. Equation 3 can thus be rewritten as:

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} p(\varphi) \prod_{a \in A} p(\mathbf{y}_a|\varphi(a)). \tag{4}$$

The above expression is further simplified by assuming that the speaking time τ_a and the interaction n-tuples $\mathbf{x}_a$ of actors a are statistically independent given the role $\varphi(a)$, thus the last equation becomes:

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} p(\varphi) \prod_{a \in A} p\left(\mathbf{x}_a|\varphi(a)\right) p(\tau_a|\varphi(a)). \tag{5}$$

The problem left open is how to estimate the different probabilities appearing in the above equation. As the components of the n-tuple $\mathbf{x}_a$ are binary, i.e. $x_{aj} = 1$ when actor a talks during segment j and 0 otherwise, the most natural way of modeling $\mathbf{x}_a$ is to use independent Bernoulli discrete distributions:

$$p(\mathbf{x}|\mu) = \prod_{j=1}^{D} \mu_j^{x_j} (1 - \mu_j)^{1-x_j}, \tag{6}$$

where D is the number of events in the SAN (see above), and $\mu = (\mu_1, \ldots, \mu_D)$ is the parameter vector of the distribution. A different Bernoulli distribution is trained for each role. The maximum likelihood estimates of the parameters μ_r for a given role r are as follows [3]:

$$\mu_{rj} = \frac{1}{|A_r|} \sum_{a \in A_r} x_{aj}, \tag{7}$$

where A_r is the set of actors playing the role r in the training set, and $\mathbf{x}_a$ is the n-tuple representing the actor a.

If the roles are independent, then $p(\varphi)$ corresponds to the following:

$$p(\varphi) = \prod_{a \in A} p(\varphi(a)) \tag{8}$$

and the a-priori probability of observing the role r can be estimated as follows:

$$p(\varphi(a)) = \frac{N_{\varphi(a)}}{N}, \tag{9}$$

Corpus	AM	SA	GT	IP	HR	WM	PM	ME	UI	ID
C1	41.2%	5.5%	34.8%	4.0%	7.1%	6.3%	N/A	N/A	N/A	N/A
C2	17.3%	10.3%	64.9%	0.0%	4.0%	1.7%	N/A	N/A	N/A	N/A
C3	N/A	N/A	N/A	N/A	N/A	N/A	36.6%	22.1%	19.8%	21.5%

Table 1. Role distribution. The table reports the percentage of time each role accounts for in C1, C2 and C3.

where N and $N_{\varphi(a)}$ are the total number of actors and the total number of actors playing role $\varphi(a)$ in the training set.

In this way, Equation 4 becomes as follows:

$$\hat{\varphi} = \arg \max_{\varphi \in \mathcal{R}^A} \prod_{a \in A} p(\mathbf{x}_a|\varphi(a))p(\tau_a|\varphi(a))p(\varphi(a)). \tag{10}$$

and the role recognition process simply consists in assigning each actor the role $\varphi(a)$ that maximizes the probability $p(\mathbf{x}_a|\varphi(a))p(\tau_a|\varphi(a))p(\varphi(a))$.

Finally, the estimation of $p(\tau|r)$ is performed using a Gaussian Distribution $\mathcal{N}(\tau|\mu_r, \sigma_r)$, where μ_r and σ_r are the sample mean and variance respectively:

$$\mu_r = \frac{1}{|A_r|} \sum_{a \in A_r} \tau_a, \tag{11}$$

$$\sigma_r = \frac{1}{|A_r|} \sum_{a \in A_r} (\tau_a - \mu_r)^2. \tag{12}$$

This corresponds to a Maximum Likelihood estimate, where a different Gaussian distribution is obtained for each role.

3.3 Experiments and Results

The experiments of this work have been performed over three different corpora referred to as C1, C2 and C3 in the following. C1 contains all news bulletins (96 in total) broadcasted by *Radio Suisse Romande* (the French speaking Swiss National broadcasting service) during February 2005. C2 contains all talk-shows (27 in total) broadcasted by *Radio Suisse Romande* during February 2005. C3 is the AMI meeting corpus [6], a collection of 138 meeting recordings involving 4 persons each and with an average length of 19 minutes and 50 seconds.

The roles of C1 and C2 share the same names and correspond to similar functions: the *Anchorman* (AM), the *Second Anchorman* (SA), the *Guest* (GT), the *Interview Participant* (IP), the *Headline Reader* (HR), and the *Weather Man* (WM). In C3, the role set is different and contains the *Project Manager* (PM), the *Marketing Expert* (ME), the *User Interface Expert* (UI), and the *Industrial Designer* (ID). See Table 1 for the distribution of roles in the corpora.

The experiments are based on a k-fold cross-validation approach ($k = 5$) [3]. The only hyperparameter to be set is the number D of segments used as events

Corpus	all (σ)	AM	SA	GT	IP	HR	WM	PM	ME	UI	ID
Automatic Speaker Segmentation											
C1	81.7 (6.9)	98.0	4.0	92.0	5.6	55.9	76.8	N/A	N/A	N/A	N/A
C2	83.2 (6.7)	75.0	88.3	91.5	N/A	29.1	9.0	N/A	N/A	N/A	N/A
C3	46.0 (24.7)	N/A	N/A	N/A	N/A	N/A	N/A	79.6	13.1	41.4	20.3
Manual Speaker Segmentation											
C1	95.1 (4.6)	100	88.5	98.3	13.9	100	97.9	N/A	N/A	N/A	N/A
C2	96.2 (2.6)	96.3	100	96.6	N/A	100	70.4	N/A	N/A	N/A	N/A
C3	51.2 (24.2)	N/A	N/A	N/A	N/A	N/A	N/A	83.3	15.9	42.0	29.0

Table 2. Role recognition performance.

in the Social Affiliation Network. At each iteration of the k-fold cross-validation, D is varied such that the value giving the highest role recognition results *over the training set* has been retained for testing. The statistical significance of performance differences is assessed with the Kolmogorov-Smirnov test [11].

The performance is measured in terms of *accuracy*, i.e. the percentage of time correctly labeled in terms of role in the test set. Each accuracy value is accompanied by the standard deviation of the accuracies achieved over the different recordings of each corpus.

The results suggest that meeting roles do not result into stable behavioral patterns (at least for what concerns the turn-taking), hence the performance on C3 is lower than the one on the other corpora. The only exception is the PM that is actually recognized to a satisfactory extent.

The performance difference when passing from manual to automatic speaker segmentation is always significant for C1 and C2 because the effectiveness of the speaker segmentation is relatively low for these corpora, thus the automatic segmentation is affected by a significant amount of errors. This results, on average, in a 10% accuracy drop.

4 Conclusions

This paper has presented two examples of how nonverbal communication can be used to understand automatically social phenomena like roles or professional activities. These works are part of a much wider range of activities in the computing community that aim at using nonverbal communication as a key towards automatic understanding of social and affective phenomena. An important effort in this sense is done by a European collaboration called Social Signal Processing Network (SSPNet). This project is building a large repository of data, tools, and publications at disposition of the scientific community. The material can be downloaded from the web portal `www.sspnet.eu` and it covers not only the problems presented in this article, but also a wide spectrum of social phenomena such as group interactions, politeness, competitive discussions, etc.

Technology of nonverbal communication, whether aimed at emotional phenomena like Affective Computing, or at social interactions like Social Signal

Processing, promises to bring significant improvement in all technologies where machines are expected to seamlessly integrate human activities, e.g. ambient intelligence, Human Computer Interaction, computer mediated communication, etc.

However, there are significant challenges that must be addressed before these improvements can actually be achieved: while the works presented in this article are based on a single modality (speech), multimodal approaches are likely to be more effective especially when the cues are ambiguous and redundancy can improve robustness. Social interaction is a inherently sequential phenomenon, but most of current approaches do not exploit human behavior dynamics because this is difficult to model. Last, but not least the integration of human sciences findings in computing technologies is not straightforward. All of these challenges open exciting research perspectives that will be addressed in the next years.

References

1. J. Ajmera. *Robust audio segmentation.* PhD thesis, Ecole Polytechnique Federale de Lausanne (EPFL), 2004.
2. C. M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.
3. C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer Verlag, 2006.
4. E. Blaauw. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication*, 14:359–375, 1994.
5. P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *in Proceedings of the Institute of Phonetic Sciences, Amsterdam*, volume 17, pages 97–110, 1993.
6. J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, et al. The AMI meeting corpus: A pre-announcement. In *Lecture notes in computer science*, volume 3869, pages 28–39, 2005.
7. K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan. *Data mining: a knowledge discovery approach.* Springer, 2007.
8. D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations from non-verbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing*, 17(3):501–513, 2009.
9. G. P.M. Laan. The contribution of intonation, segmental durations, and spectral features to perception of a spontaneous and a read speaking style. *Speech Communication*, 22:43–65, 1997.
10. J. Llisterri. Speaking style in speech research. In *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, 1992.
11. F.J. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, pages 68–78, 1951.
12. A. Vinciarelli. Capturing order in social interactions. *IEEE Signal Processing Magazine*, 26(5):133–137, 2009.
13. A. Vinciarelli and S. Favre. Broadcast news story segmentation using Social Network Analysis and Hidden Markov Models. In *Proceedings of ACM International Conference on Multimedia*, pages 261–264, 2007.

14. A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.
15. M. Weintraub, K. Taussing, K. Hunicke-Smith, and A. Snodgrass. Effect of speaking style on LVCSR performance. In *In Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 16–19, 1996.